Natural Language Processing (2025), **31**, pp. 1187–1209
doi:10.1017/nlp.2024.56

CAMBRIDGE UNIVERSITY PRESS

SQUIB

Second language learning of degree expressions: A computational approach

Yan Cong

School of Languages and Cultures, Purdue University, West Lafayette, IN 47907, USA
Email: cong4@purdue.edu

(Received 2 October 2023; revised 25 October 2024; accepted 25 October 2024; first published online 3 December 2024)

Abstract

Degrees, unlike entities or events, refer to comparative qualities and are closely tied to gradable adjectives such as "tall." Degree expressions have been explored in second language (L2) research, covering areas such as learnability, first language (L1) transfer, contrastive analysis, and acquisition difficulty. However, a computational approach to the learning of degree expressions in L2 contexts, particularly for L1 Chinese learners of English, has not been thoroughly investigated. This study aims to fill this gap by utilizing natural language processing (NLP) methods, drawing insights from recent advancements in large language models (LLMs). This study extends Cong (2024)'s general-purpose assessment pipeline to specifically analyze degree expressions, predicting that surprisal metrics will correlate with proficiency levels and distinct developmental stages of L2 learners. Crucially, we address the limitations of surprisal metrics in capturing underuse or avoidance—common in L2 development—by integrating frequency-based analyses. Using an NLP pipeline developed with Stanza, we automatically identified and analyzed degree expressions, constructing linear mixed-effects models to track L2 development trajectories. Our findings reveal that as proficiency increases, learners use complex degree expressions more frequently, supporting theories linking difficulty and learnability. Higher surprisal values are associated with lower proficiency in using degree expressions, and these surprisals are more predictive of degree expressions proficiency than classic NLP measures. These results add further evidence that LLMs and NLP tools provide valuable insights into L2 development, specifically in the domain of degree expressions, expanding upon previous research and offering new approaches for understanding L2 learning processes.

Keywords: Large language models; Degree expressions; Second language acquisition; Automatic essay scoring

1. Introduction

Degrees refer to things that can be compared, which stands in contrast with entities or events. They are closely related to gradable adjectives such as *tall*, which denotes a relation between a point on a height scale and an individual. Degree expressions have been shown to inform second language (L2) research in learnability (Lipka 2020), L1 transfer (Odlin, 1996; Lee and Li 2014; Mohri, Maruo, and Tei 2018), contrastive analysis (Broisson and Van Goethem 2018), and acquisition ease and difficulty (Yen 2003). Despite being widely examined in L2 studies, as far as we know, the learning of degree expressions have not yet been investigated computationally. Here, we attempt to provide a natural language processing (NLP) account for the L2 learning of English degree expressions in L1 Chinese learners.

In this investigation, we aim at understanding English degree expressions produced by first language (L1) Chinese learners from the perspective of NLP. The current work draws significant insight from Cong (2024), which demonstrates a systematic investigation of LLM technology,

L2 development, and automatic L2 writing assessment. Cong (2024) examines the use of *surprisals* derived from various large language models (LLMs), which refer to the negative logarithm probability of a word in a given context as calculated by an LLM, and they are widely used in computational linguistics (Michaelov and Bergen 2020; Misra 2022; Cong *et al*. 2023), speech synthesis (Kakouros *et al*. 2023), and cognitive science (Willems *et al*. 2016; Wilcox *et al*. 2018). The current work extends the general-purpose assessment pipeline in Cong (2024) to the analysis of degree expressions. We hypothesize that compared to high-proficiency L2 learners, degree expressions produced by L2 learners with low proficiency are likely to contain a high volume of unexpectedness. If LLMs are informative of L2 development in degree expressions, we predict higher surprisals to be associated with less proficient usage of degree expressions, and inversely, lower surprisals with higher proficiency in degree expressions usage. Moreover, if LLMs metrics are distinct and informative, we would expect that surprisals features lead to distinct factors and they are predictive in classifying degree expressions development stages.

Crucially, LLMs-surprisals cannot characterize underuse or avoidance, which are commonly observed in L2 development trajectories, especially for degree expressions which are hypothetically difficult to Chinese learners Eckman (1977). LLMs-surprisals are centered on the probability of encountering a word given a context in large corpora of mostly native speaker data. They measure how surprising a given input is to the model, which concerns prediction accuracy but not necessarily production patterns (Tunstall, Von Werra, and Wolf 2022). On the other hand, underuse and avoidance are production phenomena; they refer to the learner's tendency to either not use certain constructions as frequently as native speakers do (underuse) or to deliberately avoid them altogether (avoidance) (Laufer and Eliasson, 1993). These are not directly captured by a model designed to predict the likelihood of words in mostly native speaker data. This motivates us to combine both LLMs-surprisal and frequency aspects in the proposal accounting for degree expressions development. To examine the frequency proposal, we developed an NLP pipeline using Stanza (Manning *et al*. 2014; Qi *et al*. 2020) to automatically identify degree expression and its components. We examined L2 development trajectories through constructing linear mixed-effects models and calculating degree expressions usage frequencies.

Our frequency result added new empirical evidence to the research showing that difficulty and learnability are closely associated (Eckman, 1977). Further, we find that with learners' levels of proficiency advances, they produce complex degree expressions more frequently. We also find evidence that higher surprisals are associated with lower proficiency in degree expressions usage, the use of surprisals proves to be more effective than using the existing measures of cohesion in predicting proficiency of degree expressions usage, and that LLMs-surprisals capture distinct L2 degree expressions development features. These surprisals-related findings extended and added to the research by Kharkwal and Muresan (2014) and Cong (2024). Overall, our results indicate that LLMs and NLP are promising tools in improving our understanding of L2 development in degree expressions.

2. Background

2.1 Degree expressions linguistics

Our analysis is driven by the theoretical linguistic studies of degree expressions. Truth-conditionally, degree expressions such as *taller* in an utterance "Alex is taller than Kai" is interpreted as follows: there is some degree *d* such that Alex's height meets or exceeds *d* and Kai's height does not meet or exceed *d* (Heim, 1985; Neeleman, Koot, and Doetjes 2004; Kennedy 2007; Cong 2021). Descriptively, the syntax of a degree expression in English is composed of the following: target of comparison *Alex*, gradable predicate *tall*, comparative morpheme *-er*, standard marker *than*, and standard of comparison *Kai* (Stassen, 1984; Kennedy and McNally 2005). The presence and absence of these components make the syntax of degree expressions vary

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

crosslinguistically. There are two major kinds of degree expressions: phrasal and clausal comparatives (Beck, Oda, and Sugisaki 2004; Oda 2008; Sudo 2009). Consider examples in (1).

(1) a. Alex is taller than Kai. (Phrasal Comparative)
b. Alex is taller than Kai *is*. (Clausal Comparative)

Many languages including English also allow "implicit" comparatives such as (2), where there is no explicit standard marker or standard of comparison.

(2) Alex is taller. (Implicit Comparative)

(3) Kai bi Anna gao.
Kai than Anna tall
"Kai is taller than Anna" (Comparative in Mandarin Chinese)

Crosslinguistically, the comparative form of a gradable adjective *tall-er* is either derived from or identical to its plain form (also called "positive" form) *tall* (Grano 2012; Grano and Davis 2018; Cong 2021). In Chinese, on the surface, the comparative form of a gradable adjective is identical to its plain form (Grano 2012), as shown in (3). There is no morphological marking when making comparisons in Chinese. By contrast, for many other languages in the world, the comparative form is derived from the positive form (Grano and Davis 2018). For example, in English, a comparative morpheme *-er* or a word such as *more* is needed to mark comparison. As illustrated in Grano and Davis (2018), in Irish, affix *-a* marks comparison as in *arda* "taller," and in French, a word *plus* marks comparison as in *plus grand* "taller." This crosslinguistic contrast motivates us to study the learning ease and difficulty that L1 Chinese speakers experience when learning English degree expressions.

In the typology literature, Stassen (1984) and Kennedy (2007) provide comprehensive documentation about crosslinguistic variations of degree expressions. According to their documentation, Chinese is classified as an exceed comparative language, because the standard of comparison is the direct object of a transitive verb that means "exceed." Comparative-form adjectives in Chinese are of arguably equal complexity with their plain-form counterpart (Grano and Davis 2018). Different from English where both phrasal and clausal comparatives are available, Chinese only allows the standard of comparison to be phrasal (Mohri *et al*. 2018). Typologically, English is a particle comparative language: the standard of comparison is marked by a comparative particle *than*. Such typological differences in English and Chinese inspire us to investigate the role of crosslinguistic influence in degree expressions' learning process, systematically examining subtypes of degree expressions: explicit (phrasal and clausal) comparatives and implicit comparatives.

2.2 Degree expressions learning

The L2 acquisition literature on degree expressions provides various levels of insights for the learning of degree expressions. Lipka (2020) investigates the learnability of L2 English degree expressions in L1 Chinese and L1 Slavic learners. The degree expressions they examined included comparative and superlative constructions. Lipka (2020) compared the two L2 groups' syntactic awareness to a matched sample of L1 English speakers. Their findings suggested that the three groups did not perform statistically differently on items addressing degree expressions, although the L1 Chinese group gave significantly worse performance than the other groups on past tense constructions. They argued that degree expressions exist in all three languages, which explained the absence of significant difference in the learnability of degree expressions. On the other hand, they suggested that past tense exists in Slavic but not in Chinese, which explained that L1 Slavic facilitates the learnability of L2 English past tense constructions, yet L1 Chinese impedes

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

it. Relatedly, Broisson and Van Goethem (2018) examined French-speaking Belgian learners of English, indicating that discrepancy and competition between French and English degree expressions impede learning. They proposed that degree expressions learning difficulty comes from learners' effort of overcoming their preference to "more pretty" than "prettier." This is likely because learners' L1 French lacks an *-er*-like comparative morpheme. Similar crosslinguistic influence studies can be found in Mohri *et al*. (2018), which examined positive and negative L1 transfer on Japanese learners' L2 acquisition of English clausal comparatives. Along the same line of degree expressions research, Yen (2003) suggested negative L1 transfer from Chinese to the target language English. Interestingly, Yen (2003) found methodological effects that a production task was harder than a comprehension task, and for L2 learners with higher proficiency, there was less L1 transfer and methodological effect.

A crucial framework inspiring us on L1 transfer and crosslinguistic influence is the Differential Markedness Hypothesis proposed by Eckman (1977), which defined markedness as follows: a phenomenon A in some language is more marked than B if the presence of A in a language implies the presence of B but not vice versa. Eckman (1977) provides an example to illustrate the theory, demonstrating that certain languages, such as Arabic and Greek, allow passive sentences without explicitly stated agents (as shown in example 4a), but not with explicitly stated agents (as in example 4b). Conversely, languages such as English, French, and Japanese accommodate both types of passive sentences. However, there seems to be no languages where passive constructions with agents exist independently of those without agents. Thus, the existence of passive sentences with agents implies the existence of passive sentences without agents, but not vice versa. Consequently, sentences such as (4b) are considered more marked than sentences such as (4a).

(4) a. The window was closed.
    b. The window was closed by Alex.

We propose to implement the Differential Markedness Hypothesis to degree expressions. There are languages such as Chinese where only phrasal comparatives are available (Mohri *et al*. 2018), and languages like English where both clausal and phrasal comparatives are available. However, as far as our knowledge goes, there are no languages where only clausal comparatives are available. The presence of clausal comparatives implies the presence of phrasal comparatives but not the inverse. Therefore, clausal comparatives are hypothesized to be more marked than phrasal comparatives. Furthermore, Eckman (1977) hypothesizes the following: (i) areas of the target language that are different from learners' L1 and that are more marked than L1 will be difficult to learn; (ii) areas of the target language that are different from L1 but are not more marked than L1 will not be difficult to learn.

This leads us to predict that English clausal comparatives would be hard to learn for L1 Chinese learners; hence, learners may avoid using them in general. Under Eckman's framework, clausal comparatives are learning areas that are different from learners L1 and that are more marked than L1. For the other learning areas such as implicit comparatives, although the morphology of English comparatives ('*-er*'/ '*more*') is different from learners' L1 Chinese, there is no empirical evidence that this area is more marked than L1. Therefore, even though negative L1 transfer is expected, inspired by Eckman (1977) and Yen (2003), we hypothesize that learning difficulty may not necessarily persist, and such negative transfer would decrease as learners' proficiency improves.

To operate the Differential Markedness Hypothesis in the current work, we used dependency parser to identify degree expression and its subtypes. These subtypes are based on crosslinguistic degree semantics and typology studies (Stassen, 1984; Kennedy and McNally 2005; Kennedy 2007; Cong 2021), as discussed in the previous subsection. We argue that LLMs metrics are informative only when there are sufficient input representations for LLMs to abstract patterns from. Therefore, we complement LLMs metrics with investigations of frequency. We acknowledge that including a

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

dependency parser might lead to a rather complex pipeline. However, we maintain that this inclusion is necessary, since degree expression is sophisticated, understudied, and there is avoidance and underuse among L2 learners. As far as our knowledge goes, there is no published engineering tool in how to automatically identify degree expressions. Our utilization of dependency parsing can bridge the gap.

2.3 LLM-surprisals and L2 studies

Previous work on L2 studies utilize LLMs and NLP tools for L2-related tasks, such as automatic comprehension questions generation, essay readability assessment, conversational systems, speech processing, grammatical error detection and correction, feedback generation, annotation, bilingual learning modeling, and so on (Bommasani *et al*. 2021; Keim and Littman 2022; Reyes *et al*. 2022; Alic *et al*. Alic *et al*. 2022; Bexte, Horbach, and Zesch 2022; Takano and Ichikawa 2022; Han *et al*. 2023; Koraishi 2023; Cong 2024). We derive significant insights from previous studies, highlighting the growing interest in utilizing LLMs in L2 research. For example, Han *et al*. (2023) investigates interactions between students and LLMs, particularly how students engage with ChatGPT to revise their essays. This study employs the RECIPE4U corpus, which includes conversation logs, students' intents, self-rated satisfaction, and their essay edit histories. By analyzing these components, the research aims to assess ChatGPT's effectiveness in enhancing EFL (English as a Foreign Language) writing education. Relatedly, Koraishi (2023) explores the diverse applications of ChatGPT as a tool for EFL teachers, focusing on material development and assessment. It illustrates how ChatGPT can streamline the creation of engaging and contextually relevant resources tailored to individual learners' needs. Additionally, it discusses the use of ChatGPT in text assessment, providing real-time, personalized feedback to improve learners' performance and overall learning experience.

Against the background of LLMs technology and L2 studies, a critical NLP innovation in the current work is the introduction of LLM-surprisals in the measurement of L2 degree expressions. Even before the advent of LLMs, surprisal has been analyzed in L2 research and particularly in measuring L2 writing proficiency. For example, Kharkwal and Muresan (2014) employed surprisal as a predictor of essay quality among Swedish students learning English. Their study provided evidence supporting the hypothesis that lower proficiency L2 learners' writings tend to contain a higher volume of unexpected elements compared to those of higher proficiency learners. In another study by Kaan and Chun (2018), the prime verb surprisal effects were examined by comparing the transitive production preferences of native American English speakers and Korean L2 learners of English. Their investigation was built up on the view that when the prime structure is unexpected given the preceding words in the sentence, surprisal is presumably high and that surprisal can serve as a valid predictor of processing difficulty (Hale 2001; Smith and Levy 2008). Surprisals, as a good metric for prediction error, have been examined to understand the L2 processing mechanisms. Higher prime surprisal is a reflection of higher prediction error, and it is presumably associated with a larger priming effect, which is perhaps part of an adaptive response facilitating communication efficiency (Fine and Florian Jaeger 2013). Based on previous work about surprisals and adaptation, in Kaan and Chun (2018), the written priming study investigated the use of double object and prepositional phrase datives. Overall, this line of research showed that both L1 and L2 Korean groups exhibited cumulative adaptation effects for both types of datives, indicating that L1 and L2 speakers employ similar processing mechanisms. Additionally, they maintained that differences in adaptation can be attributed to the relative frequency of these structures in their respective languages. This line of work inspires us to connect LLMs-surprisals with degree expressions learnability.

Relatedly, in light of recent progress in LLMs techniques, Cong (2024) attempts to investigate the application of LLM models in L2 automated essay scoring and discuss their benefits and drawbacks. By incorporating indices of lexical diversity and syntactic complexity from

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

previous research, Cong (2024) connects past findings with future directions in L2 research. By comparing LLMs-surprisals with the existing NLP indices, Cong (2024) highlights significant advancements in the application of LLMs-surprisals to L2 development. Cong (2024)'s comparative approach between LLMs-surprisals and traditional linguistic indices provides a robust framework for understanding the potential of LLMs in capturing the complexity of L2 learning processes. Building on Cong (2024), our proposed research aims to extend previous methods by specifically targeting the acquisition of degree expressions, a critical yet under-explored area in L2 development. Through incorporating machine learning classifiers and factor analyses, we seek to further delineate the efficacy of LLMs-surprisals in predicting L2 learners' performance compared to the established indices. The methodological enhancement in LLMs not only promises a more granular understanding of how L2 learners internalize degree expressions but also contributes to refining assessment tools in L2 education. We are hopeful that LLMs-surprisals can inform L2 degree expressions development research.

2.4 Present study

The current work is significantly inspired by and meanwhile different from Cong (2024) in the the following ways. Cong (2024) focused on LLMs-surprisals and provided a general NLP pipeline that is not tailored for analyzing specific constructions. The current work integrates the surprisal and frequency aspects into the analysis of L2 degree expressions. While LLMs-surprisals can provide useful information about the predictability of linguistic constructions in native language corpora, they are not equipped to directly capture phenomena such as underuse or avoidance in L2 development. These phenomena require a more nuanced analysis of learner language production and its deviations from native speaker patterns, which goes beyond the scope of what surprisal measures can offer. Despite that advanced frameworks and techniques such as instruction tuning and direct preference optimization have presented LLMs with both positive and negative evidence (Brown *et al*. 2020; Ouyang *et al*. 2022; Rafailov *et al*. 2024), and next-word prediction tasks also (indirectly) shape the model's understanding of linguistic constraints, we maintain that by complementing LLMs-surprisal with frequency analyses tailored for underuse and avoidance, we can capture L2 development more effectively. LLMs-surprisals do not account for the absence of certain constructions in learner data. If a learner avoids using complex degree expressions, the surprisal model would not necessarily flag this absence since it operates on what is present in the input, not what is missing. Analyzing underuse and avoidance requires comparing the frequency and variety of constructions in learner language with those in native speaker benchmarks, which LLMs-surprisals are not designed to do.

Further, LLMs lack the insight into learner strategies. Underuse and avoidance are often strategic choices by learners to circumvent linguistic challenges (Laufer and Eliasson, 1993). These strategies can be influenced by factors such as proficiency level, confidence, and exposure to the target language. LLMs-surprisals do not provide insight into the cognitive or strategic processes behind language production decisions. They simply reflect how unexpected a construction is based on large-scale native language usage. Last but not the least, we need to consider the corpus-based nature of LLMs-surprisals. LLMs are trained on extensive corpora of mostly native speaker data, and LLMs-surprisal values reflect patterns in this data. However, these patterns may not align with the interlanguage of L2 learners, who have different exposure and developmental paths. Overall, these considerations motivated us to identify underuse or avoidance in learner language through frequency analysis of learner corpora, which can highlight discrepancies between the expected natural proficient usage and actual learner production.

The current work expands Cong (2024) in LLMs' efficacy evaluation through including more classic NLP indices. Different from Cong (2024), the current work compared LLMs-surprisals

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

and Tool for the Automatic Analysis of Cohesion (TAACO 2.0.4, Crossley *et al*. (2019)) indices specifically in the context of degree expressions. This cohesion-related comparison is motivated by the following considerations. First, comparing efficacy of surprisals and the existing cohesion indices can provide enhanced objectivity and consistency in assessing degree expressions. LLMs can analyze degree expressions by evaluating the statistical predictability of words and structures within these expressions. TAACO provides indices that measure various aspects of textual cohesion, including how effectively degree expressions and comparatives are made and linked in a text. These indices can reflect the clarity and coherence of degree expressions, essential for evaluating L2 learners' proficiency in using such structures. Comparing these two approaches allows us to explore the extent to which statistical language predictability (surprisal) aligns with or complements classic cohesion measures. This can help identify which method or combination of methods offers a more reliable and objective assessment.

Second, such comparison of cohesion-related efficacy can provide a comprehensive evaluation of degree expression mastery. LLMs-surprisals focus on the probabilistic aspects of language use, providing insights into the fluency and naturalness of degree expressions (Cong 2024). This can help identify whether learners are using comparative structures fluently and naturally. On the other hand, TAACO cohesion indices focus on the structural and rhetorical aspects of writing, providing insights into how well degree expressions contribute to the cohesion of the text. We assume that effective use of degree expressions enhances textual clarity and comparison. By comparing these methods, we can achieve a more holistic understanding of L2 learners' mastery of degree expressions, integrating both surface-level fluency and deeper cohesion. This comprehensive evaluation can also lead to more constructive feedback for learners.

Third, cohesion-related efficacy comparison provides empirical validation and has practical implications. As argued in Cong (2024), LLMs-surprisals require empirical validation to ensure their efficacy and reliability in the context of degree expressions. Comparing them with TAACO indices, which have been widely studied and validated, can provide critical insights into their practical applicability. Although TAACO cohesion indices are robust, they may benefit from the inclusion of LLMs-surprisal measures to capture additional dimensions of proficiency in using degree expressions. It is also worth noting that the current work built predictive classifiers to provide a quantifiable measure of indices efficacy. By comparing the performance of classifiers built on different sets of features (e.g., LLMs indices *vs*. classic complexity indices), we can determine which set of indices provides better predictive power and thus is more useful for degree expression evaluation. Such efficacy comparison can not only empirically validate each approach's strengths and limitations in assessing degree expressions but also lead to practical recommendations for their use in educational settings. This will ultimately enhance L2 instruction and evaluation in the use of degree expressions, highlighting the potential of integrating cutting-edge AI with established tools, making advanced assessment techniques for degree expressions more accessible and practical for everyday educational use.

Last but not the least, although Cong (2024) showed evidence that LLMs-surprisals differ from the tool for the automatic analysis of syntactic sophistication and complexity (TAASSC) (Lu 2010; Kyle 2016) in efficacy of indexing L2 development and proficiency, and that combing the two can improve efficacy, the comparison between LLMs-surprisals and TAASSC indices in Cong (2024) concerns efficacy. It did not examine whether the two are characterizing different L2 constructs conceptually; and if they are, what aspects of L2 degree expressions they are characterizing. To address those questions, the current work conducted factor analysis, extracting key components that TAASSC and LLMs-surprisals are potentially capturing. Identifying key constructs or dimensions that contribute to degree expression proficiency is crucial for understanding L2 development in degree expressions. Overall, by exploring how these approaches complement and enhance each other conceptually and quantitatively, we can develop more reliable and interpretable tools for L2 assessment in degree expressions, ultimately benefiting learners, educators, and researchers in the field of applied linguistics.

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Table 1. A subset of PELIC that we extracted for our degree expressions analysis

Level	Learners total	Sentences total	Cmp total
2	2	86	4
3	44	4642	195
4	76	9081	282
5	48	6897	235

3. Method

3.1 Degree expressions dataset

Following Cong (2024), we used the publicly available University of Pittsburgh English Language Institute Corpus (PELIC) (Juffs, Han, and Naismith 2020; Naismith, Han, and Juffs 2022), a learner corpus of written texts, containing 4.2 million words. This collection, gathered over a period of seven years within the University of Pittsburgh's Intensive English Program, includes writings from over 1,100 students with varied linguistic backgrounds and levels of proficiency. Unlike many learner corpora that are cross-sectional, PELIC is longitudinal, providing enhanced opportunities to monitor progress in a real classroom environment. We extracted degree expressions from PELIC (Table 1), where "Learners total" is the total number of learners whose paragraph writing contains Cmp sentences, namely Cmp learners. "Sentences total" is the total number of sentences produced by the Cmp learners. "Cmp total" is the total number of sentences that include comparative degree expressions, among "Sentences total."

We followed Cong (2024) to focus on level 3,4,5 learners and removed level 2 in our investigation, since there were significantly less data for level 2 learners. Moreover, we find that sentences containing Cmp are not frequently used in L2 paragraph writing. This leads us to normalize Cmp subtypes usage by the total Cmp sentences, rather than the total number of sentences, since there are very few Cmp subtype degree expressions in a learner's writing.

3.2 Degree expressions extraction

To extract degree expressions, we used Stanza dependency parser as illustrated in Figure 1. We identified explicit (phrasal and clausal comparatives) and implicit comparatives. Specifically, we used nominal inflectional features *"Degree=Cmp"* to extract sentences that contain comparative degree expressions (Cmp), including phrasal comparatives (Phrasal_Cmp) such as example (1a) "Alex is taller than Kai," clausal comparatives (Clausal_Cmp) such (1b) "Alex is taller than Kai is," and implicit comparatives (Implicit_Cmp) such as example (2) "Alex is taller."

To separate two types of explicit comparatives Clausal_Cmp and Phrasal_Cmp, we proposed Semgrex rules such as the following: {pos: JJR}> /ccomp|advcl/ {} for cases like "Alex did it better than she did" and {pos: RB}> advmod {pos:RBR}> /ccomp|advcl/ {} for "Alex marched more quickly than she thought." To improve parsing accuracy, before applying the Semgrex rules, we conducted a Corpus of Linguistic Acceptability (CoLA) classification task using the T5-large LLM (Raffel *et al*. 2020). CoLA is a linguistic acceptability NLP task where a model checks if the sentence is grammatically acceptable given a text prompt. Only the acceptable sentences underwent Semgrex parsing. After identifying explicit comparatives (Clausal_Cmp and the Phrasal_Cmp), the rest of the degree expressions in the large Cmp pool were annotated as Implicit_Cmp, for instance, "Alex felt better."

In addition to parsing the syntax of degree expressions into explicit and implicit comparatives, the same dataset was parsed based on their morphology. Sentences in the Cmp pool containing

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Table 2. Subtypes of degree expressions extracted from the L2 corpus

Classification basis	Subtype	Example	NLP operation	Variable name
syntax	explicit - phrasal	*Alex is taller than Kai*.	dependency parse	Phrasal_Cmp
	explicit - clausal	*Alex is taller than Kai is*.	CoLA, Semgrex	Clausal_Cmp
	implicit	*Alex is taller*.	dependency parse	Implicit_Cmp
morphology	*more/less*	*This is more expensive than that*.	string matching	More_Cmp
	ER	*This is longer than that*.	string matching	ER_Cmp

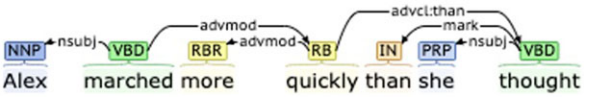

Figure 1. Dependency parse of an English clausal comparative sentence "*Alex marched more quickly than she thought*." with the Stanza dependency parser.

"more" or "less" were classified as More_Cmp. On the other hand, for sentences containing morphemes such as "*-er*", they were classified as ER_Cmp. To sum up, our degree expressions extraction NLP pipeline classifies and outputs the following subtypes of degree expressions as illustrated in Table 2.

Table 2 illustrated all the subtypes of degree expressions analyzed in the current work. The classification is based on the theoretical linguistics literature about degree expressions. Table 2 also listed key examples, operation steps, and variable names for each subtype.

3.3 Degree expressions analysis

Frequency analysis was conducted after extracting degree expressions sentences from the corpus, including the raw counts as well as percentage of each degree expression subtype. For surprisals, we calculated LLMs-surprisals at the target word of each tokenized comparative sentence. The "target word" refers to the word that has the Stanza-based comparative degree feature Cmp. Following Misra (2022) and Cong (2024), the surprisal for the target in the context was computed as (1). When w_t was tokenized into multiple subword tokens, we used the average of the subword tokens probabilities. We did not measure other metrics such as maximum or summation, because we intended to control and normalize the length for subword tokens, such that the influences of target word's tokenization length on surprisals get reduced. We additionally computed surprisals at the level of the entire sentence. The surprisal of the sentence is the summation of the surprisal scores of each token, normalized by the sentence length:

$$Surprisal(w_t) = -logP(w_t|w_{1\ldots t-1}) \quad (1)$$

We experimented with three variations of LLMs using the decoder-transformer architecture, which includes the decoder part of transformer and generates text by predicting the next word in a sequence. The models varied in size to assess how scaling impacts performance: GPT-2 with 124 million parameters (Radford *et al.* 2019), DistilGPT-2 with 82 million parameters (Sanh *et al.* 2019), and GPTNeo with 1.3 billion parameters (Gao *et al.* 2020; Black *et al.* 2022). Additionally, we included the GPT-3 model text-davinci-002 (OpenAI 2023) and the open-source decoder-only transformer LLaMA2 with 7 billion parameters (Touvron *et al.* 2023). We utilized minicons (Misra 2022), an open-source tool that offers a standard API (application programming interface) for conducting systematically behavioral analyses of LLMs. All the selected LLMs, except for

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

"text-davinci-002, are available on HuggingFace (https://huggingface.co/). OpenAI's davinci-002 is a paid model (https://openai.com/blog/openai-api).

Further, we conducted a comparative analysis of efficacy in the LLM-surprisals and the classic NLP indices of cohesion and complexity. First, we used TAACO (version 2.0.4, Crossley *et al.* (2019)) to build predictive classifiers. Operationally, TAACO is a cohesion measurement tool, which calculates 150 indices. TAACO reported on a variety of validated local and global features of cohesion (Crossley *et al.* 2019), which methodologically aligns with LLMs-surprisals of degree expressions at the target word and the sentence level. TAACO indices can be classified into three types: (i) type-token ratio indices (ttr, including specific parts of speech, lemmas, bigrams, trigrams and more), (ii) adjacent overlap indices (at both the sentence and paragraph level), and (iii) connectives indices. For type (i), we examined the number of unique lemmas (types) divided by the number of total running lemmas (tokens) ("lemma_ttr"), and Number of unique bigram lemmas (types) divided by the number of total bigram lemmas (tokens) ("bigram_lemma_ttr"). For type (ii), we did not include any TAACO indices because our measure unit is one sentence instead of a paragraph; thus, there is no adjacent sentence pairs for which we could use the average latent semantic analysis cosine similarity ("lsa_1_all_sent"). For type (iii), we calculated number of basic connectives divided by number of words in text ("basic_connectives"), number of sentence linking words divided by number of words in text ("sentence_linking"), and number of all connectives divided by number of words in text ("all_connective").

Besides cohesion indices, in a factor analysis, we included the widely studied indices of complexities (Kyle and Crossley 2018; Kyle 2021). We used TAASSC (Lu 2010; Kyle 2016), an accessible tool for syntactic analysis, assessing various indices concerning syntactic development. It includes traditional measures of syntactic intricacy such as the average length of T-units (a T-unit consists of an independent clause and any dependent clauses attached to it), as well as detailed measures of phrasal (such as the ratio of adjectives to noun phrases), noun phrase complexity (such as adjectival modifiers per nominal), and clausal complexity (such as the density of adverbials per clause). Furthermore, it incorporates indices rooted in usage-based theories of language acquisition, which rely on the frequency distributions of verb argument constructions. A full index description spreadsheet can be found in https://www.linguisticanalysistools.org/taassc.html. Operationally, for TAASSC, we computed all types of the indices, including (1) clause complexity, (2) noun phrase complexity, and (3) syntactic sophistication. Note that in our factor analysis, the clause complexity and noun phrase complexity types of indices generated almost all zeros in the output file. This was possibly due to the fact that the input is at sentence level, which can be too short for TAASSC to generate meaningful indices such as number of dependents per indirect object or number dependents per nominal complement. As a consequence, our factor analysis was mostly consisted of syntactic sophistication.

4. Results

4.1 L2 development in degree expressions

4.1.1 Surprisal aspects

First, we examined learners' L2 development in degree expressions through surprisals measurement. Visually and numerically examining the data suggests that its distribution does not meet assumptions of parametric statistical tests such as *t*-test for dependent samples. Descriptive statistics for the LLMs variables were given in Table 3.

We reproduced Cong (2024)'s surprisal method in degree expressions. We conducted a non-parametric statistical hypothesis test: the independent two-samples Wilcoxon signed-rank test, to compare LLMs-surprisals of degree expressions produced by learners with adjacent proficiency levels, and to estimate whether the population means ranks differ statistically. Findings were visualized in Figure 2. The alpha level in this work is 0.05. Effect sizes of statistical tests in Figure 2 were listed in Table 4.

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Table 3. Descriptive statistics for the LLMs variables

Variable	Mean	SD	Min	Max
gpt2_target_surprisal	7.04	3.88	0.00	22.18
distilgpt2_target_surprisal	7.48	4.00	0.00	21.79
gptneo_target_surprisal	6.83	4.02	0.00	21.79
davinci002_target_surprisal	4.39	2.68	0.01	14.74
llama2_target_surprisal	6.66	3.93	0.00	21.39
gpt2_sentence_surprisal	4.63	0.96	2.39	9.32
distilgpt2_sentence_surprisal	5.06	1.01	2.67	10.07
gptneo_sentence_surprisal	4.40	0.99	2.01	9.66
davinci002_sentence_surprisal	4.61	0.99	1.86	9.56
llama2_sentence_surprisal	4.40	0.90	2.34	9.28

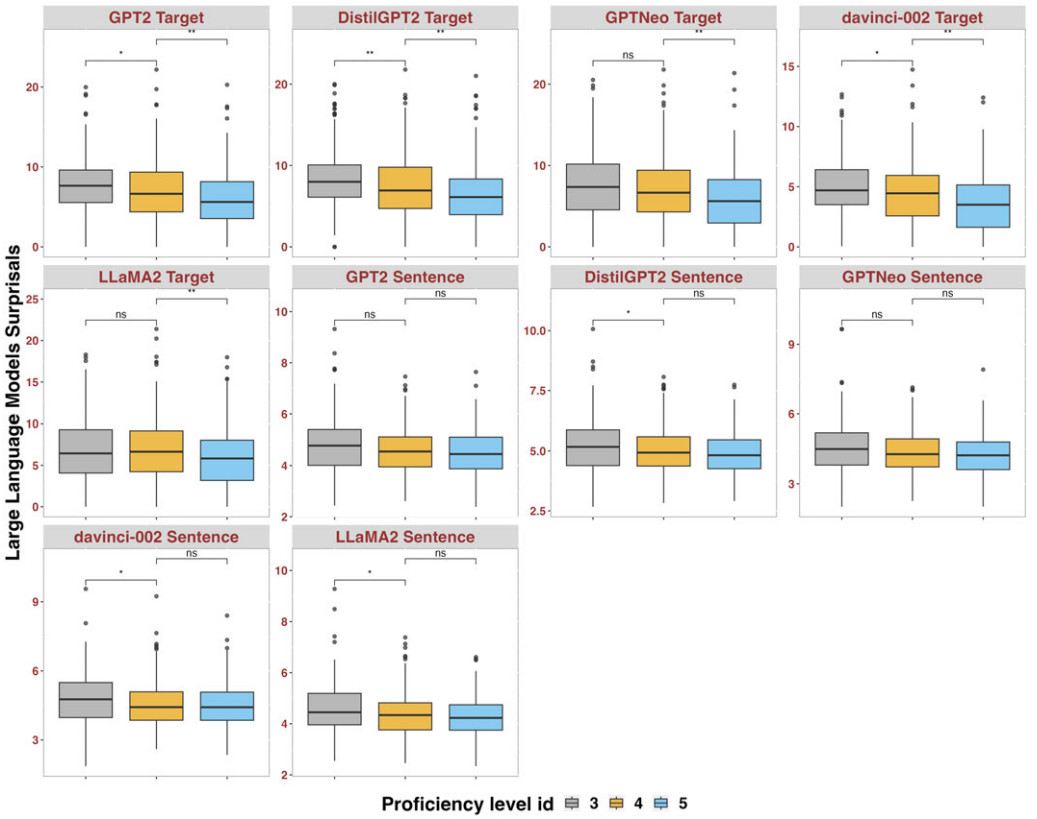

Figure 2. L2 development in degree expressions indexed by LLMs-surprisals. $^{*}p<0.05$; $^{**}p<0.001$; $^{***}p<0.0001$; *ns*, no significant difference.

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Table 4. Effect sizes of paired comparisons

Variable	Level 3 vs. Level 4	Level 4 vs. Level 5
gpt2_target_surprisal	0.125*	0.124**
distilgpt2_target_surprisal	0.153**	0.123**
gptneo_target_surprisal	0.072	0.136**
davinci002_target_surprisal	0.11*	0.147**
llama2_target_surprisal	0.012	0.012**
gpt2_sentence_surprisal	0.099	0.053
distilgpt2_sentence_surprisal	0.115*	0.049
gptneo_sentence_surprisal	0.097	0.045
davinci002_sentence_surprisal	0.106*	0.02
llama2_sentence_surprisal	0.116*	0.03

Significance notation: $^{*}p < 0.05$; $^{**}p < 0.001$; $^{***}p < 0.0001$.

In Figure 2, across LLMs, surprisals of degree expressions at the target and sentence level showed a decrease as L2 learners' proficiency increases. This is as predicted and aligns with previous L2 findings that LLMs-surprisals decrease as their training and bilingual system progress (Kharkwal and Muresan 2014; Cong 2024). Our results extended this method and finding to the understudied area of degree expressions learning. Concretely, at the target level, statistical group differences were found in GPT2, DistilGPT2, and davinci-002 surprisals across all the adjacent proficiency levels. Interestingly, larger LLMs such as GPTNeo and LLaMA2 did not show significantly more sensitivity. They were only able to distinguish level 4 from level 5 but not level 3 from level 4. The sentence level surprisal metrics, on the other hand, suggested fewer significance. DistilGPT2, davinci-002, and LLaMA2 sentence surprisals can distinguish degree expressions usage proficiency between level 3 and 4 but not between level 4 and 5. GPT2 and GPTNeo did not show evidence that they can benchmark proficiency levels in degree expression usage. Small effect sizes were found for all the paired comparisons, as illustrated in Table 4.

4.1.2 Frequency aspects

To understand the developmental trajectory of degree expressions' learning, we further analyzed L2 learners' usage frequency at various proficiency levels. For each degree expression subtype, we fitted a linear mixed-effects model using the raw counts of a degree expression subtype as the dependent variables. The independent variable is the proficiency level (level 3, 4, 5, as provided by the L2 corpus PELIC). We took participant ID (anon_id) as a random intercept in our models. We used the LME4 package (Kuznetsova, Brockhoff, and Christensen 2017) for model fitting. Results were summarized in Table 5.

Table 5 showed the estimates, standard errors (*SE*), *t* values, and *p* values for the fixed effect of level_id. The results indicated that there is a significant decrease of phrasal comparatives usage from level 3 to level 4, and a significant decrease of "*-er*" comparatives usage from level 3 to level 4. In contrast, level 4 learners use significantly more implicit ("Alex felt better") and "more"/ "less" comparatives than level 3 learners do. From the aspect of syntax, we speculate that an increase of implicit comparatives might be an indicator of flexibility in degree expressions usage as learners' interlanguage system advances. From the aspect of morphology, we speculate that more complex vocabularies are being used as learners' mastery of degree expressions improves; hence, "more"

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Table 5. Linear mixed-effects models summary for each degree expression subtype

Variable	contrast	Estimate	*SE*	*t*	*p*
Intercept		0.41	0.04	9.74	<2e-16
Phrasal_Cmp	Level 3 - Level 4	0.20	0.05	3.91	0.00
	Level 4 - Level 5	−0.003	0.05	−0.07	0.998
Intercept		1.179e-02	7.996e-03	1.47	0.14
Clausal_Cmp	Level 3 - Level 4	0.004	0.01	0.33	0.94
	Level 4 - Level 5	−0.004	0.01	−0.42	0.91
Intercept		0.59	0.04	14.05	<2e-16
Implicit_Cmp	Level 3 - Level 4	−0.2	0.05	−3.91	0.00
	Level 4 - Level 5	0.003	0.05	0.07	0.998
Intercept		0.42	0.04	9.37	5.98e-16
More_Cmp	Level 3 - Level 4	−0.15	0.06	−2.77	0.02
	Level 4 - Level 5	0.03	0.05	0.62	0.81
Intercept		0.58	0.04	13.19	<2e-16
ER_Cmp	Level 3 - Level 4	0.15	0.06	2.77	0.02
	Level 4 - Level 5	−0.03	0.05	−0.62	0.81

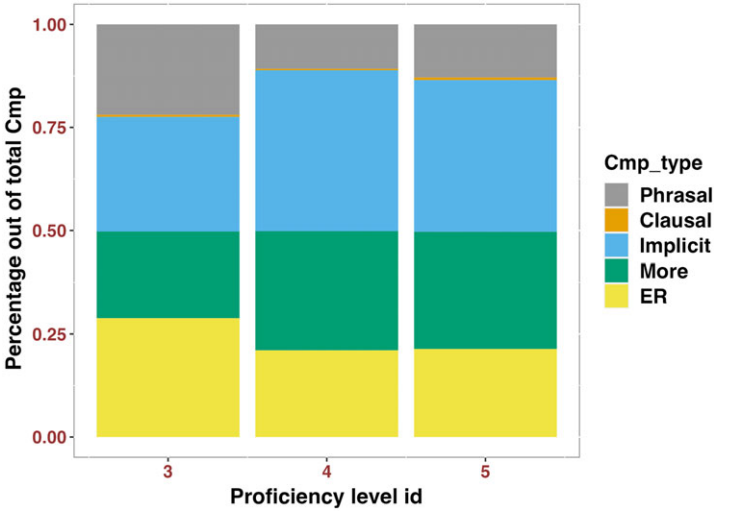

Figure 3. Degree expressions subtypes usage development across proficiency levels.

is needed to modify complex multi-syllabic adjectives. Further, no significant usage change were found from level 4 to level 5 across degree expressions subtypes, suggesting that degree expressions learnability manifests at a relatively early learning stage, and it becomes less evident as learners arrive at the advanced stage. We found random effects for all the four aforementioned models, suggesting that there are individual differences in the usage of degree expressions. We additionally visualized the usage percentage of each degree expression subtype in a stacked bar plot (Figure 3).

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Table 6. Evaluation metrics of the random forest models predicting writing proficiency

Feature selection	Accuracy	Precision	Recall	F1-Score	AUC(SD)
Existing indices of cohesion	0.60	0.52	0.52	0.52	0.53(0.06)
LLMs-surprisals	0.60	0.55	0.55	0.55	0.58(0.08)
LLMs-surprisals and existing indices of cohesion	0.65	0.58	0.58	0.58	0.61(0.08)

Aside from the shrinkage of explicit phrasal comparatives and *-er* comparatives usage and the corresponding expansion of implicit comparatives and *more/less* comparatives, Figure 3 illustrated that, as predicted, clausal comparatives are rarely used across L2 learners' interlanguage development stages. Our finding also added new empirical evidence to the research on the Differential Markedness Hypothesis (Eckman, 1977). To L1 Chinese learners of English, clausal comparatives are different from their L1 and they are marked, which increases learning difficulty and impedes the learnability. Translated into L2 usage, our frequency results showed that learners are inclined to strategically avoid using this marked degree expression that is not readily available in their L1.

4.2 Efficacy comparison in degree expressions assessment

To investigate how LLMs advance automatic assessment of L2 degree expressions proficiency, taking the established classic NLP indices as comparison baseline, we built predictive models to classify L2 essays into high versus low writing proficiency. It is worth noting that the degree expressions dataset is a subset of the L2 corpus. The proficiency scores provided by PELIC are for the overall (writing) proficiency but not a particular expression. In other words, there is no gold standard labels provided by human professionals assessing degree expressions proficiency. Therefore, we took Writing_Sample (overall essay score rated by domain experts) as the gold standard label and as a proxy of degree expressions proficiency, and we made the assumption that a good essay score (a high Writing_Sample score) is an indicator of learners' general mastery of different language phenomena in the target language, possibly including degree expressions.

4.2.1 Cohesion indices

First, we compared efficacy of LLMs-surprisals with the classic cohesion indices. Specifically, the two proficiency classes were set based on the median of the variable Writing_Sample. Data with a Writing_Sample score 3.3 or higher is labeled as high degree expressions proficiency otherwise as low. We constructed three binary random forest classifiers using scikit-learn (Pedregosa *et al.* 2011): a classifier with LLMs derived surprisals, one with the selected existing indices of cohesion as calculated by TAACO, and one with the combination of both. Optimal features were identified using recursive feature elimination. Hyperparameters used in the random forest models include the following: number of trees = 100, the function to measure the quality of a split=entropy, and do not split subsets smaller than 2. Random forest models were evaluated using repeated stratified K-fold cross-validation, the number of times the cross-validator needs to be repeated is 3, and the number of folds is 10. The best random forest classifier combines both LLMs-surprisals and the existing cohesion indices (accuracy 0.65, mean AUC (area under the curve) 0.61 (SD = 0.08)). Metrics for each random forest model are reported in Table 6.

Our findings suggested that LLMs derived indices are better than the existing cohesion indices in predicting L2 degree expressions proficiency, although the difference is marginal. A random forest model with the existing indices showed at-chance performance, whereas the same model with LLMs metrics showed above-chance performance. Adding existing indices on top of LLMs

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Table 7. Factor analysis of the surprisals and existing (syntactic) sophistication indices

Factor 1	Loading	Factor 2	Loading	Factor 3	Loading
news_av_freq_log	0.92	llama2_sentence_surprisal	0.76	gpt2_target_surprisal	0.73
acad_av_freq_log	0.92	distilgpt2_sentence_surprisal	0.75	distilgpt2_target_surprisal	0.71
mag_av_freq_log	0.91	gptneo_sentence_surprisal	0.75	gptneo_target_surprisal	0.71
all_av_freq_log	0.90	gpt2_sentence_surprisal	0.72	llama2_target_surprisal	0.60
mag_av_freq_type	0.86	gptneo_target_surprisal	0.61	distilgpt2_len_tokens	0.45
fic_av_freq	0.86	gpt2_target_surprisal	0.59	gptneo_len_tokens	0.45
all_av_lemma_freq_log	0.85	distilgpt2_target_surprisal	0.54	MLS	0.44
news_av_freq	0.85	llama2_target_surprisal	0.48		
acad_av_freq_type	0.81	MLT	−0.47		
vac_frequency	0.74	gptneo_len_tokens	−0.72		
frequency	0.68	distilgpt2_len_tokens	−0.72		
diversity_and_frequency	0.62	MLS	−0.73		
acad_av_lemma_freq	0.56				
acad_av_lemma_freq_type	0.56				

measures did not significantly improve the classifier's performance. Overall, results indicated that there is potential in utilizing LLMs to improve cohesion metrics' automatic assessment accuracy.

For the random forest model where both LLMs and classic NLP indices were included, the top five decisive features are all LLMs metrics. Ranked in descending order based on importance score: target surprisal as calculated by DistilGPT2, GPT2, and LLaMA2, sentence surprisals as calculated by LLaMA2, and target surprisals as calculated by GPTNeo. The existing cohesion indices did not enter the top 10 decisive features set. This indicated that LLMs showed larger effects than the existing indices in degree expressions proficiency classification. Further, this rank also suggested that larger LLMs such as LLaMA2 and GPTNeo showed efficacy, but they may not necessarily give better performance than the smaller ones in detecting L2 learners' degree expressions proficiency.

4.2.2 Complexity indices

A factor analysis was conducted on TAASSC measures of syntactic complexity (Kyle 2016). We used *FactorAnalysis* from scikit-learn (Pedregosa *et al.* 2011), with varimax rotation. Table 7 showed loading for each factor. Only features that are strongly correlated with the factor object were listed. Factors loading were given in Table 7.

Three factors were identified, illustrated in Table 7. Factor 1 reflected aspects of syntactic complexity and sophistication: news_av_freq_log represents average construction frequency log-transformed, in reference to the newspaper sub-corpora in COCA (Corpus of Contemporary American English). Similarly, acad_av_freq_log, mag_av_freq_log, and all_av_freq_log are in reference to the academic, magazine, and all written sub-corpora in COCA, respectively. mag_av_freq_type represents average constructions frequency (types only) in reference to the magazine sub-corpora. Similarly for acad_av_freq_type, which refers to the academic sub-corpora. fic_av_freq represents average construction frequency in reference to the fiction sub-corpora. Similarly for news_av_freq, which is in reference to the newspaper sub-corpora,

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

all_av_lemma_freq_log represents average lemma frequency log-transformed, in reference to the all written sub-corpora. vac_frequency represents VAC (verb-argument construction) frequency and direct objects component. And frequency represents frequency component. Overall, these measures indicate how frequently certain syntactic constructions and lexical items occur in different sub-corpora of COCA. The inclusion of these indices in factor 1 suggests that L2 learners' ability to use degree expressions effectively is tied to their familiarity with constructions that are common in various contexts (e.g., newspapers, academic texts, magazines). Factor 1 reveals about syntactic complexity: higher frequencies of complex constructions in learner texts indicate a more advanced grasp of syntactic structures necessary for forming degree expressions. For instance, learners who can use less common constructions typical of academic or fiction texts might demonstrate higher proficiency. Moreover, the log-transformed frequency measures reflect learners' exposure to and use of sophisticated syntactic forms. This sophistication is crucial for correctly constructing and understanding degree expressions, which involve intricate syntax.

Factor 2 concerns sentence-level LLMs-surprisals and text length: MLT represents mean length of T-unit, namely number of words in text divided by number of T-units in text. MLS represents mean length of sentence. These measures provide insights into the overall fluency of learners' writing. Longer T-units and sentences suggest that learners are capable of producing longer, more complex, and elaborate sentences, which is indicative of advanced language proficiency. len_tokens refer to the total number of tokens in the sequence, as calculated by an LLM. Specifically, longer sentence lengths (hence longer len_tokens) and T-units can indicate greater fluency, suggesting that learners are comfortable constructing extended comparative statements such as degree expressions without breaking them into simpler, shorter sentences. From the aspect of complexity, factor 2 reveals that the ability to maintain coherence and grammaticality in longer sentences shows a higher level of syntactic complexity, essential for expressing nuanced comparisons accurately. Factor 3 concerns target word surprisals and text length. This factor concerns particularly word usage in degree expressions. It suggested that target surprisals make unique contributions and can stand-alone as a distinct factor.

To sum up, factor analyses suggested that the surprisals and most of the existing syntactic complexity indices capture different aspects of degree expressions proficiency. LLMs-surprisals were associated with sequence length. Interestingly, for factor 2, LLMs-surprisals showed positive loading, whereas the length features showed the inverse, indicating that longer sequence may be associated with lower degree expressions surprisals.

4.2.3 Alternative surprisal indices

In addition to utilizing LLM-based surprisal predictors, we compared these with the surprisal indices proposed by Kharkwal and Muresan (2014), which employed probability context-free grammar (PCFG) surprisal. According to their work, surprisal approaches zero when a word is highly predictable in a given context. Kharkwal and Muresan leveraged surprisal to measure the complexity of sentence processing. They computed surprisal using a broad-coverage top-down parser (Roark *et al*. 2009), which determines the negative log probability of a word based on its preceding context through prefix probabilities. This method quantifies the unexpectedness of text sequences given their sentential context. Their findings indicated that average surprisal values in essays decrease with EFL training and that there is an inverse relationship between surprisal and essay scores. Inspired by the methodology of Kharkwal and Muresan (2014), we calculated PCFG-surprisals and analyzed the correlation between LLM-based surprisal indices and PCFG surprisal indices with the writing scores (Writing_Sample) of L2 learners. Spearman's correlations coefficients and *p* values are reported in Table 8.

Our results replicated previous findings that LLMs-surprisal values decrease as learners' proficiency increases. This suggests that as learners become more proficient, their degree expressions usage becomes more predictable according to the LLMs, reflecting greater fluency and

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Table 8. Spearman's correlations between Writing_Sample and different kinds of surprisals

Surprisal	Coefficent	*p*
gpt2_target_surprisal	−0.184	<0.0001
distilgpt2_target_surprisal	−0.185	<0.0001
gptneo_target_surprisal	−0.16	0.0001
davinci002_target_surprisal	−0.181	<0.0001
llama2_target_surprisal	−0.163	<0.0001
gpt2_sentence_surprisal	−0.137	0.001
distilgpt2_sentence_surprisal	−0.159	0.0001
gptneo_sentence_surprisal	−0.126	0.003
davinci002_sentence_surprisal	−0.165	<0.0001
llama2_sentence_surprisal	−0.127	0.002
PCFG_target_surprisal	0.055	0.161
PCFG_sentence_surprisal	−0.034	0.381

natural usage of degree expressions. Additionally, we found that the two PCFG surprisal features did not show significant correlations with learners' writing proficiency, whereas all the LLMs-surprisals did. Also, compared to PCFG-surprisals, LLM-based surprisal indices showed stronger correlation coefficients, indicating their potential strength in advancing L2 research.

5. Discussions

5.1 L2 development

The frequency analysis is informative of degree expressions' L2 developmental trajectory over three proficiency levels. Our results suggested that as L2 learners' interlanguage system develops incrementally, they tend to produce more "*more*"/ "*less*" comparatives. This implies that with learners' interlanguage system evolving, they learn to correctly spell out the standard of the comparison, even when the comparison construction's syntax is complex and different from that in their L1. This supports Yen (2003)'s hypothesis that negative L1 transfer tends to disappear as learners' proficiency increases. Aside from adding supportive evidence to the existing research, we discover new insights. Our findings (see section 4.1.2) reveal that compared to earlier stage, later stage L2 learners are more and more inclined to produce more "*more*"/ "*less*" comparatives than "*-er*" comparatives. It is likely that at the earlier stage, L2 learners mostly produce simple monosyllabic words, which are compatible with "*-er*" comparatives. As their interlanguage knowledge accumulates and grows, L2 learners produce complex gradable adjectives that require modifications such as "*more*" or "s*less*." Further, our findings about significant underuse of clausal comparatives can be accounted for by Eckman (1977)'s theory. The Differential Markedness Hypothesis explains the difficulty for L1 Chinese learners to learn English clausal comparatives, which are different and more marked than learners' L1. Such underuse and avoidance are also likely caused by crosslinguistic influence. Clausal comparatives are not readily available in learners' native language Chinese, which impedes the learning. This finding echoes previous studies by Lipka (2020) and Broisson and Van Goethem (2018).

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

For surprisals analysis, our findings (see section 4.1.1 and 4.2) add to the research showing that in general, LLMs-surprisals decrease as learners' interlanguage system progresses (Kharkwal and Muresan 2014; Cong 2024). We further generalized this finding to the learnability of degree expressions. The factor analysis involving TAASSC complexity indices and LLMs-surprisals provided a multifaceted view of L2 development in degree expressions. We speculate that degree expressions complexity can be decomposed into three major components. Specifically, Factor 1 (syntactic complexity and sophistication) measures are related to construction frequency in various COCA sub-corpora. This factor only concerns TAASSC classic indices, indicating that LLMs-surprisals reveal dimensions in degree expressions learnability that are distinctive from the classic sophistication indices. Factor 2 (sentence-level LLMs-surprisals and text length) include a mixture of surprisals and the classic fluency measures such as MLT and MLS. This factor also implies that LLMs-surprisals appear to overlap and is closely associated with sequence length. Factor 3 (target word surprisals and text length) is a second factor suggesting that LLMs-surprisals and the classic indices of sequence length are closely clustered. Together, these factors reveal that mastering degree expressions in an L2 involves both an understanding of complex syntactic patterns and the ability to use them fluently and naturally.

The comparisons between LLMs- and PCFG-surprisals (see section 4.2.3) also have implications for L2 development understanding. As discussed in Cong (2024), LLM-based surprisals indices can provide insights into the intuitive grasp of language by L2 learners, highlighting their ability to use degree expressions naturally and fluently. This aligns with the broader goal of achieving naturalness and "native-like" proficiency in L2 learning. On the other hand, our findings suggested that PCFG-surprisals, while also valuable, primarily reflect the syntactic complexity and processing demands of sentences. They are effective for measuring how learners handle complex grammatical structures, which may offer a complementary perspective to the more context-sensitive naturalness-based LLMs measures.

5.2 LLMs' efficacy

First, the random forest classifier showed an above-chance efficacy when integrating LLMs generated features, whereas the accuracy was at chance level when using only the existing indices. We thus speculate that LLMs can improve automatic proficiency assessments of degree expressions. We also found that surprisal as calculated by DistilGPT2 was the most decisive feature. Interestingly, features generated by bigger LLMs such as LLaMA2 or GPTNeo were not as informative. This supports previous findings that depending on areas, scaling is not always beneficial, and that compressing LLMs can achieve better performance on downstream tasks (Rae *et al*. 2021; Li *et al*. 2021; Cong 2024). Overall, our predictive classifiers results provide further evidence that LLMs can enhance and advance L2 proficiency studies.

Second, factor analyses detailed that LLMs metrics differ from most of the existing proficiency assessment indices. This suggests that the underlying mechanisms of LLMs in differentiating proficiency levels are distinct from classic indices such as cohesion, syntactic complexity, and sophistication. This comparison contributes to the understanding of LLMs' efficacy in language assessment, in a sense that it reveals LLMs' understanding of interlanguage development as as a dynamic system, which goes above and beyond cohesion and sophistication. The trajectory of L2 acquisition of degree expressions is not only attributed to classic indices, such as cohesion, complexity, and sophistication, but it is also a reflection of unexpectedness, adding to the research in Cong (2024).

Further, LLM-based surprisal indices exhibited stronger correlation coefficients compared to PCFG-surprisals, providing further evidence that LLM-based measures may be more effective in capturing the intricacies of L2 degree expressions proficiency. LLMs' better performance might be attributed to sensitivity to context. LLM-based surprisal indices benefit from the extensive training data and advanced algorithms of language models, enabling them to better understand and

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

predict word usage in various contexts. This results in more accurate and nuanced surprisal measurements. In contrast, PCFG-surprisals, while effective, may be limited by the constraints of their grammatical frameworks and less extensive training data. The stronger correlations observed with LLM-based surprisal indices highlight their potential for advancing L2 assessment. These indices can provide more detailed feedback on learners' proficiency, particularly in complex linguistic constructs such as degree expressions.

5.3 Real-world applications

Our results provide an approach to reveal L2 learners' interlanguage system's development in degree expressions. L2 teachers and researchers may benefit from specifying and personalizing the design in teaching materials, based on learners L1 and proficiency levels. L2 learners may benefit from consciously producing clausal comparatives or other constructions that are absent or different in their L1. Our findings would improve learners' awareness of crosslinguistic contrasts. Our approach adds to the applications of accurate assessment methods, which are crucial for diagnosing learner difficulties, providing targeted feedback, and guiding instructional strategies.

Following and expanding Cong (2024), we used one of the many published LLMs-related APIs called minicons (Misra 2022), which enables consistent and streamlined approach to compute surprisals from multiple LLMs. Our degree expression extraction is also built on top of one of the many existing parsers such as Stanza Manning *et al*. (2014). Overall, we hope to build pipelines utilizing the validated libraries and APIs, such that they are reliable and computationally light. We did not investigate how fine-tuning LLMs would influence automatic assessment of degree expressions. We acknowledge that fine-tuning or training LLMs could lead to a model without using dependency parsers, and we leave it for future investigation. In the same line of investigation as in Cong (2024), we intend to showcase an approach for future L2 researchers to take *pretrained* LLMs off-the-shelf and use their laptops to collect linguistically meaningful LLMs measures without wandering in the massive LLMs zoo or tediously normalizing various LLMs outputs. Without ample supply of computation power, L2 researchers can still probe LLMs using our proposed approach and gain quantifiable information from pretrained LLMs. The use of LLMs offers scalability and adaptability, making them suitable for a wide range of educational and research applications.

5.4 Limitations and future research

We acknowledge that there are limitations in our current study. First, L2 learning of degree expressions is relatively understudied, and our NLP pipeline calls for cross-validations and independent reproduction. We made a GitHub repository publicly available, where we provided python script for extracting degree expressions and calculating surprisals, aside from the metadata. We hope this can facilitate research reproducibility.

Second, for the current study, we did not filter the prompt in the writing tasks, given the consideration that degree expressions in general have a very low frequency in PELIC L1 Chinese learners' writing and further filtering might lead to an even smaller dataset. We acknowledge that certain prompts or questions might facilitate learners to produce degree expressions, while others might not. For future research, we hope to carefully control prompts in our NLP pipelines and expand our datasets by including new open-source corpora besides PELIC.

Further, we did not collect human professionals' proficiency scores particularly for degree expressions. We used the overall essay score provided by PELIC (Writing_Sample). We acknowledge that the assumption of associating overall writing proficiency with degree expressions usage proficiency needs some refinement. As discussed, degree expressions are not frequently produced in L2 learners' essays. Such underuse and avoidance can be caused by not only crosslinguistic influence but also limited exposure. It is likely that degree expressions are of low frequency in L1

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

English speakers' essays, requesting new explanations as to whether this will lead to the underuse of degree expressions in L2. We leave these for future research to address.

6. Conclusion

To computationally quantify degree expressions usage development in L2 learning, we used frequency and LLMs-surprisal scores. Our findings are aligned with previous results that lower surprisals are associated with proficient and coherent texts (Kharkwal and Muresan 2014; Cong 2024). Our method and findings add to the research showing that surprisal scores are promising metrics. They have potential to enhance L2 development research and automatic writing assessment. Leveraging LLMs in education contexts also speaks to recent works by Crossley and Holmes (2023); Crossley *et al*. (2023). Further, we examined and discussed how different LLMs unveil the trajectory of L2 development in degree expressions. We provided interpretations on what this reveals about the models' understanding of language cohesion, complexity, and general development. We hope our study will inspire more refined research on the use of NLP in studying degree expressions and surprisals in texts produced by L2 language learners.

Data availability. The second language data were drawn from PELIC https://github.com/ELI-Data-Mining-Group/PELIC-dataset. Corpus citation: Juffs, A., Han, N-R., and Naismith, B. (2020). The University of Pittsburgh English Language Corpus (PELIC) [Data set]. https://zenodo.org/records/4577423. Software and metadata are available in this GitHub repository: https://github.com/yancong222/NLPDegreeExpression.

Acknowledgment. We would like to thank Emmanuele Chersoni for inspiring us to utilize LLMs-surprisals, Brian Buccola for insightful discussions on degree expressions, and Phillip Wolff for informing us on transformer architectures and dependency and constituency parsers. Personal communication with the aforementioned scholars helped streamline the methodology. We acknowledge Cameron Pilla's help with optimizing the machine learning pipeline. We are also grateful for the interesting discussions with Gaurav Kharkwal and Brian Roark. This project is supported by School of Languages and Cultures, College of Liberal Arts, Purdue University. All errors remain mine.

Funding statement. This project is funded by the College of Liberal Arts, Purdue University.

Competing interests. The authors declare no competing interests.

References

Alic S., Demszky D., Mancenido Z., Liu J., Hill H. and Jurafsky D. (2022). Computationally identifying funneling and focusing questions in classroom discourse. In Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applicatinos (BEA 2022), pp. 224–233.

Beck S., Oda T. and Sugisaki K. (2004). Parametric variation in the semantics of comparison: Japanese vs English. *Journal of East Asian Linguistics* **13**(4), 289–344.

Bexte M., Horbach A. and Zesch T. (2022). Similarity-based content scoring-how to make S-BERT keep up with BERT. In Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pp. 118–123.

Black S., Biderman S., Hallahan E., Anthony Q., Gao L., Golding L., He H., Leahy C., McDonell K. and Phang J. (2022). Gpt-neox-20b: An open-source autoregressive language model. In Proceedings of the Workshop on Challenges & Perspectives in Creating Large Language Models, pp. 95–136.

Bommasani R., Hudson D. A., Adeli E., Altman R., Arora S., von Arx S., Bernstein M. S., BohgJ., Bosselut A. and Brunskill E. (2021). On the opportunities and risks of foundation models. Preprint at arXiv https://doi.org/10.48550/arXiv.2108.07258

Broisson Z. and Van Goethem K. (2018). Comparative constructions in french-speaking belgian learners of english: a contrastive approach. *CECL Papers,* pp. 26–28.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G. and Askell A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901.

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Cong Y. (2021). Competition in natural language meaning: the case of adjective constructions in mandarin chinese and beyond. *Doctoral dissertation*. Michigan State University.

Cong Y. (2024). Demystifying large language models in second language development research. *Computer Speech & Language* **89**, 101700.

Cong Y., Chersoni E., Hsu Y.-Y. and Blache P. (2023). Investigating the effect of discourse connectives on transformer surprisal: language models understand connectives; even so they are surprised. In Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP. EMNLP 2023, pp. 222–232.

Crossley S., Choi J. S., Scherber Y. and Lucka M. (2023). Using large language models to develop readability formulas for educational settings. In International Conference on Artificial Intelligence in Education, Springer, pp. 422–427.

Crossley S. and Holmes L. (2023). Assessing receptive vocabulary using state-of-the-art natural language processing techniques. *Journal of Second Language Studies* **6**(1), 1–28.

Crossley S. A., Kyle K. and Dascalu M. (2019). The tool for the automatic analysis of cohesion 2.0: integrating semantic similarity and text overlap. *Behavior Research Methods* **51**, 14–27.

Eckman F. R. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning* **27**, 315–330.

Fine A. B. and Florian Jaeger T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science* **37**, 578–591.

Gao L., Biderman S., Black S., Golding L., Hoppe T., Foster C., Phang J., He H., Thite A. and Nabeshima N. (2020). The pile: An 800gb dataset of diverse text for language modeling. Preprint at arXiv https://doi.org/10.48550/arXiv.2101.00027

Grano T. (2012). Mandarin *hen* and universal markedness in gradable adjectives. *Natural Language & Linguistic Theory* **30**, 513–565.

Grano T. and Davis S. (2018). Universal markedness in gradable adjectives revisited: the morpho-semantics of the positive form in arabic. *Natural Language & Linguistic Theory* **36**, 131–147.

Hale J. (2001). A probabilistic earley parser as a psycholinguistic model. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, Association for Computational Linguistics, pp. 1–8.

Han J., Yoo H., Myung J., Kim M., Lee T. Y., Ahn S.-Y., Oh A. and Answer A. N. (2023). Exploring student-chatgpt dialogue in efl writing education. In Proceedings of the 37th NeurIPS Workshop on Generative AI for Education (GAIED), pp. 1–15.

Heim I. (1985). *Notes On Comparatives and Related Matters*. Austin: Ms., University of Texas, pp. 24–43.

Juffs A., Han N. and Naismith B. (2020). The University of Pittsburgh English Language Institute Corpus (PELIC) [Data set]. Available at http://doi.org/10.5281/zenodo.3991977 (accessed July 2021).

Kaan E. and Chun E. (2018). Priming and adaptation in native speakers and second-language learners. *Bilingualism: Language and Cognition* **21**, 228–242.

Kakouros S., Šimko J., Vainio M. and Suni A. (2023). Investigating the utility of surprisal from large language models for speech synthesis prosody. Preprint at arXiv https://doi.org/10.48550/arXiv.2306.09814

Keim G. and Littman M. (2022). Selecting context clozes for lightweight reading compliance. In Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pp. 167–172

Kennedy C. (2007). Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* **30**, 1–45.

Kennedy C. and McNally L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language* **81**, 345–381.

Kharkwal G. and Muresan S. (2014). Surprisal as a predictor of essay quality. In Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014), pp. 54–60

Koraishi O. (2023). Teaching english in the age of ai: embracing chatgpt to optimize efl materials and assessment. *Language Education and Technology (LET Journal)*, **3**(1), 55–72.

Kuznetsova A., Brockhoff P. B. and Christensen R. H. (2017). LmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* **82**(13), 1–26.

Kyle K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. Doctoral dissertation. Georgia State University. https://doi.org/10.1111/modl.12468

Kyle K. (2021). Natural language processing for learner corpus research. *International Journal of Learner Corpus Research*, **7**(1), 1–16.

Kyle K. and Crossley S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal* **102**(2), 333–349.

Laufer B. and Eliasson S. (1993). What causes avoidance in l2 learning: L1-l2 difference, l1-l2 similarity, or l2 complexity? *Studies in Second Language Acquisition* **15**(1), 35–48.

Lee S. J. and Li X. (2014). The acquisition of comparative constructions by english learners of chinese: an explorative study from a college chinese language classroom. *Chinese as a Second Language Research* **3**(1), 53–78.

Li T., Mesbahi Y. E., Kobyzev I., Rashid A., Mahmud A., Anchuri N., Hajimolahoseini H., Liu Y. and Rezagholizadeh M. (2021). A short study on compressing decoder-based language models. Preprint at arXiv https://doi.org/10.48550/arXiv.2110.08460

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Lipka O. (2020). Syntactic awareness skills in english among children who speak slavic or chinese languages as a first language and english as a second language. *International Journal of Bilingualism* **24**(2), 115–128.

Lu X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* **15**(4), 474–496.

Manning C. D., Surdeanu M., Bauer J., Finkel J. R., Bethard S. and McClosky D. (2014). The stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60.

Michaelov J. A. and Bergen B. K. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? Preprint at arXiv https://doi.org/10.48550/arXiv.2010.04844

Misra K. (2022). Enabling flexible behavioral and representational analyses of transformer language models. Preprint at arXiv https://doi.org/10.48550/arXiv.2203.13112

Mohri F., Maruo K. and Tei R. (2018). L2 research on clausal comparatives in english: positive and negative l1-transfer. *Fukuoka University Review of Literature & Humanities* **49**(4), 1001–1018.

Naismith B., Han N.-R. and Juffs A. (2022). The university of pittsburgh english language institute corpus (PELIC). *International Journal of Learner Corpus Research* **8**(1), 121–138.

Neeleman A., Koot H.v d and Doetjes J. (2004). Degree Expressions. *The Linguistic Review*, **21**(1), 1–66. https://doi.org/10.1515/tlir.2004.001

Oda T. (2008). Degree constructions in Japanese. *Doctoral dissertation*. University of Connecticut. Available from ProQuest Dissertations & Theses Global Closed Collection; ProQuest One Academic. (304641792). https://www.proquest.com/dissertations-theses/degree-constructions-japanese/docview/304641792/se-2

Odlin T. (1996). On the recognition of transfer errors. *Language Awareness* **5**(3-4), 166–178.

OpenAI. (2023). OpenAI GPT-3 API [text-davinci-003]. Available at https://platform.openai.com/docs/models

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A. and et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35**, 27730–27744.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O. and et al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12**, 2825–2830.

Qi P., Zhang Y., Zhang Y., Bolton J. and Manning C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. Preprint at arXiv https://doi.org/10.48550/arXiv.2003.07082

Radford A., Wu J., Child R., Luan D., Amodei D. and Sutskever I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog* **1**(8), 9.

Rae J. W., Borgeaud S., Cai T., Millican K., Hoffmann J., Song F., Aslanides J., Henderson S., Ring R., Young S. (2021). Scaling language models: Methods, analysis & insights from training gopher. Preprint at arXiv https://doi.org/10.48550/arXiv.2112.11446

Rafailov R., Sharma A., Mitchell E., Manning C. D., Ermon S. and Finn C. (2024). Direct preference optimization: your language model is secretly a reward model. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Preprint at arXiv https://doi.org/10.48550/arXiv.2305.18290

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W. and Liu P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* **21**(1), 5485–5551.

Reyes L. L. A., Ibañez M. A., Sapinit R., Hussien M. and Imperial J. M. (2022). A baseline readability model for Cebuano. Preprint at arXiv https://doi.org/10.48550/arXiv.2203.17225

Roark B., Bachrach A., Cardenas C. and Pallier C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In Proceedings of the Conference on empirical methods in natural language processing (EMNLP), pp. 324–333

Sanh V., Debut L., Chaumond J. and Wolf T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. Preprint at arXiv https://doi.org/10.48550/arXiv.1910.01108

Smith N. J. and Levy R. (2008). Optimal processing times in reading: A formal model and empirical investigation. In Proceedings of the Annual Meeting of the Cognitive Science Society, **30**, pp 595–600.

Stassen L. (1984). The comparative compared. *Journal of Semantics* **3**(1-2), 143–182.

Sudo Y. (2009). Invisible degree nominals in Japanese clausal comparatives. In Proceedings of the 5th Workshop on Altaic in Formal Linguistics, Cambridge, Mass: MITWPL, pp. 285–295.

Takano S. and Ichikawa O. (2022). Automatic scoring of short answers using justification cues estimated by BERT. In Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pp. 8–13

Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., Bashlykov N., Batra S., Bhargava P. and Bhosale S. (2023). Llama 2: Open foundation and fine-tuned chat models. Preprint at arXiv https://doi.org/10.48550/arXiv.2307.09288

Tunstall L., Von Werra L. and Wolf T. (2022). *Natural Language Processing with Transformers*. O'Reilly Media, Inc.

https://doi.org/10.1017/nlp.2024.56 Published online by Cambridge University Press

Wilcox E., **Levy R.**, **Morita T. and Futrell R.** (2018). What do RNN language models learn about filler-gap dependencies? Preprint at arXiv https://doi.org/10.48550/arXiv.1809.00042

Willems R. M., **Frank S. L.**, **Nijhof A. D.**, **Hagoort P. and Van den Bosch A.** (2016). Prediction during natural language comprehension. *Cerebral Cortex* **26**(6), 2506–2516.

Yen R.-P. (2003). Second Language Acquisition of English Comparative Constructions: A Case Study of Senior High School Students in Taiwan. Master Thesis, National Taiwan Normal University. Available at https://www.airitilibrary.com/Article/Detail/U0021-2603200719134046

Cite this article: Cong Y (2025). Second language learning of degree expressions: A computational approach. *Natural Language Processing* **31**, 1187–1209. https://doi.org/10.1017/nlp.2024.56